第27卷 第4期 工 程 数 学 学 报 Vol. 27 No. 4
2010年08月 CHINESE JOURNAL OF ENGINEERING MATHEMATICS Aug. 2010

文章编号:1005-3085(2010)04-0757-04

同一语言在不同进制下的正则性研究

孙旭清， 吴庚申

(青岛远洋船员学院机电系，青岛 266071)

摘 要: 本文对同一语言在不同进制的表示下正则性的问题进行了研究，证明了当p与q互质的时候，存在这样的语言L，使得L在p进制表示下是正则语言 (字母表为$\{0,1,\cdots,p-1\}$)，但是在q表示下是非正则语言 (字母表为$\{0,1,\cdots,q-1\}$)。而且p与q互质这一条件是必要的。

关键词: 正则语言；进制；泵引理

分类号: **AMS**(2000) 68Q45 **中图分类号**: TP301.1 **文献标识码**: A

有穷自动机[1-3]是一种非常简单，但是又非常重要的计算模型，它所对应的语言—正则语言也可以通过正则表达式来定义，正则表达式是一个非常强大的工具，例如正则表达式$[a-z0-9]^+$可以匹配任意的由字符或者数字组成的字符串。正则表达式广泛的应用在UNIX命令中，例如grep。正则表达式还可以应用于词法分析，模式匹配等领域。

在本文中我们将对正则语言的如下数学性质进行研究：同一语言在不同的进制表示下 (对应不同的字符表)，其正则性是否保持。正则性在很多数学变换下都是保持的，例如集合的交、并、补运算等[2]。在本文中我们将证明，在改变数的进制的时候，语言的正则性并不保持。

本文的结构如下：首先我们给出一些文章中需要的定义和引理，然后我们证明将一个语言由3进制变为2进制，正则性并不保持，即定理1。之后我们将这一结论推广到任意的互质的两个进制p和q (定理2)。

定义1 对于一个由自然数构成的语言L，我们用$(L)_p$表示语言L在p进制下的表示，即将L中的每一个元素都表示为p进制数，相应的字母表$\Sigma=\{0,1,\cdots,p-1\}$。我们把$(L)_{10}$也简记作$L$。

*例如*由全体偶数构成的语言$EVEN=(EVEN)_{10}=\{2,4,6,8,\cdots\}$，而$(EVEN)_2=\{10,100,110,1000,\cdots\}$。由全体素数构成的语言$PRIME=(PRIME)_{10}=\{2,3,5,7,11,\cdots\}$，而$(PRIME)_2=\{10,11,101,111,1011,\cdots\}$。

引理1 设L是一个正则语言，则存在常数n (与L有关)，使得对于任何L中的字符串w，如果w的长度$|w|\geq n$，那么w就可以被分成3个子串，$w=xyz$，满足：

1) $|y|>0$，即y不是空串；
2) $|xy|\leq n$；
3) 对所有的$k\geq 0$，字符串$xy^kz\in L$。

定理1 存在一个语言L，它在3进制下的语言$(L)_3$是正则语言，但是在2进制下的语言$(L)_2$是非正则语言。

证明 考虑语言$L=\{3^n\,|\,n\geq 0\}$，L的最初几项分别是：$1,3,9,27,\cdots$

$$(L)_3=\{1,10,100,1000,\cdots\},\quad (L)_2=\{1,11,1001,11011,\cdots\}.$$

收稿日期: 2008-04-24. **作者简介**: 孙旭清 (1952年4月生)，男，高级工程师. 研究方向：船舶电气及自动化.

我们将证明 $(L)_3$ 是正则语言，但 $(L)_2$ 不是正则语言。

一个接受 $(L)_3$ 的确定性有穷自动机 A：A 共有两个状态 $Q=\{q_0,q_1\}$，q_0 是初始态，q_1 是接受态。$\delta(q_0,1)=q_1$，$\delta(q_1,0)=q_1$。

下面我们证明 $(L)_2$ 不是正则语言。假设 $(L)_2$ 是正则语言，则存在着满足引理 1 中的常数 n。取 $m>n$，因为 $3^m\in L$，而显然 3^m 在 2 进制表示中的位数 $\geq m+1>n$ (因为其在 3 进制表示中有 $m+1$ 位)，所以根据引理 1，3^m 在 2 进制表示下可以被分成 3 段，$3^m=(xyz)_2$，满足引理 1 的 3 条性质。设 3 部分的长度分别是 a, b, c，由 $3^m=(xyz)_2$ 我们有

$$3^m=(x)_2\cdot 2^{b+c}+(y)_2\cdot 2^c+(z)_2. \tag{1}$$

由引理 1 中的 2) 我们知道

$$a+b=|xy|\leq n,$$

所以

$$c\geq m+1-n\geq 2.$$

另一方面由引理 1 中的 3) 我们知道对所有的 $k\geq 0$, $(xy^kz)_2\in(L)_2$，特别的 $(xyyz)_2\in(L)_2$，因此存在正整数 s，使得 $(xyyz)_2=3^s$，所以

$$3^s=(x)_2\cdot 2^{2b+c}+(y)_2\cdot 2^{b+c}+(y)_2\cdot 2^c+(z)_2. \tag{2}$$

(1), (2) 两式相减得

$$3^s-3^m=(x)_2\cdot\left(2^{2b+c}-2^{b+c}\right)+(y)_2\cdot 2^{b+c}=2^{b+c}\left[(x)_2(2^b-1)+(y)_2\right].$$

显然 $s>m$，所以 $3^m|3^s-3^m$，故由上式知

$$3^m\,|\,2^{b+c}\left[(x)_2(2^b-1)+(y)_2\right].$$

因为 $\gcd(2,3)=1$，所以 $\gcd(2^{b+c},3^m)=1$。因而

$$3^m\,|\,(x)_2(2^b-1)+(y)_2. \tag{3}$$

由引理 1 中的 1) 我们知道 $b>0$，因此

$$(x)_2(2^b-1)+(y)_2\geq(x)_2>0.$$

结合 (3) 可知

$$(x)_2(2^b-1)+(y)_2\geq 3^m,$$

因而

$$(x)_2\cdot 2^{b+c}+(y)_2\cdot 2^c+(z)_2\geq 2^c\left[(x)_2\cdot 2^b+(y)_2\right]\geq 4\cdot 3^m. \tag{4}$$

最后一步我们使用了 $c\geq 2$ 这一条件。综合 (1), (4) 两式可得 $3^m\geq 4\cdot 3^m$，矛盾。所以假设不成立，即 $(L)_2$ 不是正则语言。 **证毕**

事实上，定理 1 可以推广到任意满足 $\gcd(p,q)=1$ 的两个进制 p 和 q。

定理 2 设 p, $q>1$ 满足 $\gcd(p,q)=1$，则存在一个语言 L，它在 p 进制下的语言 $(L)_p$ 是正则语言，但是在 q 进制下的语言 $(L)_q$ 是非正则语言。

证明 考虑语言 $L=\{p^n\,|\,n\geq 0\}$，易知 $(L)_p$ 是正则语言，下面证明 $(L)_q$ 是非正则语言。假设 $(L)_q$ 是正则语言，则存在着满足引理 1 中的常数 n。取充分大的 $m>n$，使得 p^m 在 q 进制

表示中的位数 $> n$。因为 $p^m \in L$，所以根据引理1，p^m 在 q 进制表示下可以被分成3段，$p^m = (xyz)_q$。设3部分的长度分别是 a, b, c，则

$$p^m = (x)_q \cdot q^{b+c} + (y)_q \cdot q^c + (z)_q. \tag{5}$$

由于 $a + b = |xy| \leq n$，所以 $c \geq m + 1 - n \geq 2$。另一方面，由引理1 $(xyyz)_q \in (L)_q$，因此存在正整数 s，使得 $(xyyz)_q = p^s$，所以

$$p^s = (x)_q \cdot q^{2b+c} + (y)_q \cdot q^{b+c} + (y)_q \cdot q^c + (z)_q. \tag{6}$$

(6) 式减去 (5) 得

$$p^s - p^m = (x)_q \cdot \left(q^{2b+c} - q^{b+c} \right) + (y)_q \cdot q^{b+c} = q^{b+c}\left[(x)_q(q^b - 1) + (y)_q \right].$$

显然 $s > m$，故由上式知

$$p^m \mid q^{b+c}\left[(x)_q(q^b - 1) + (y)_q \right].$$

因为 $\gcd(p, q) = 1$，所以

$$p^m \mid (x)_q(q^b - 1) + (y)_q. \tag{7}$$

因为 $b = |y| > 0$，因此

$$(x)_q(q^b - 1) + (y)_q \geq (x)_q > 0.$$

故由 (7) 可知

$$(x)_q(q^b - 1) + (y)_q \geq p^m,$$

因而

$$(x)_q \cdot q^{b+c} + (y)_q \cdot q^c + (z)_q \geq q^c\left[(x)_q \cdot q^b + (y)_q \right] \geq q \cdot p^m. \tag{8}$$

综合 (5), (8) 两式可得 $p^m \geq q \cdot p^m$，矛盾。所以假设不成立，即 $(L)_q$ 不是正则语言。 **证毕**

定理2中 $\gcd(p, q) = 1$ 这一条件是必要的，去掉这一条件后结论不成立。例如：$p = 2$, $q = 4$，可以证明对于一个语言 L，如果 $(L)_2$ 是正则的，那么 $(L)_4$ 一定是正则的。

定理3 设 $p, d > 1$，对于语言 L，它在 p 进制下的语言 $(L)_p$ 是正则语言，当且仅当它在 p^d 进制下的语言 $(L)_{p^d}$ 是正则语言。

引理2 设 L 是字母表 Σ 上的正则语言，映射 $h : \Sigma \to T$，则字母表 T 上的语言 $h(L)$ 也是正则语言。

引理3 设 L 是字母表 T 上的正则语言，映射 $h : \Sigma \to T$，则字母表 Σ 上的语言 $h^{-1}(L)$ 也是正则语言。

注 $h^{-1}(L)$ 定义为 $h^{-1}(L) = \{w \mid h(w) \in L\}$。

引理2，3的证明可以在参考文献 [2] 中找到。

定理3的证明 取

$$\Sigma = \{0, 1, \cdots, p^d - 1\}, \quad T = \{0, 1, \cdots, p - 1\}^d = \{(0, \cdots, 0), \cdots, (p - 1, \cdots, p - 1)\}.$$

定义映射 $h : \Sigma \to T$, $h(x) = x$ 在 p 进制下的表示。

假设语言 L 在 p^d 进制下的语言 $(L)_{p^d}$ 是正则语言，根据引理3，语言 $h((L)_{p^d})$ 是正则语言，即语言 $(L)_p$ 是正则语言。反之假设语言 L 在 p 进制下的语言 $(L)_p$ 是正则语言，根据引理2，语言 $h^{-1}((L)_p)$ 是正则语言，即语言 $(L)_{p^d}$ 是正则语言。 **证毕**

参考文献:

[1] Du D Z, Ko Ker I. Problem Solving in Automata, Languages, and Complexity[M]. New York: Wiley-Interscience, 2001
[2] Hopcroft J, Motwani R, Ullman J. Introduction to Automata Theory, Languages, and Computation[M]. Massachusetts: Addison-Wesley, 2000
[3] Sipser M. Introduction to the Theory of Computation[M]. Boston: PWS Publishing, 1997

The Regularity of a Language under Different Bases

SUN Xu-qing, WU Geng-shen

(Qingdao Ocean Shipping Mariners College, Qingdao 266071)

Abstract: The change of the regularity of a language under different bases is investigated in this paper. We show that for any different bases p and q, if $\gcd(p,q) = 1$, then there exists a language L, such that under base p (with alphabet $\{0, 1, \cdots, p-1\}$) L is a regular language, but under base q (with alphabet $\{0, 1, \cdots, q-1\}$) L is not a regular language.

Keywords: regular language; number system; pumping lemma

Received: 24 Apr 2008. **Accepted:** 13 Sep 2009.